A4 Corpus annotation

Sayaka Nameki

4.1 Introduction

- Anotationはmark upと密接に関わってくる

Anotationとは

- コーパスの構文解析と品詞タグ付けをコード化すること
- 「書き言葉または話し言葉の電子コーパスデータに解釈や言語情報を付け加えること」
- Mark-Upは比較的検証可能な情報

  ↔Annotationは解釈言語情報

つまり、人間のテキスト理解

例えば曖昧な言葉はmark-upではなくannotationで証明される

4.1 Introduction

この章の流れ

1. Corpus anotationの長所と短所
2. Corpus annotation で最もよく使われるタイプ
3. 独立型コーパスについて

4.2 Anotationの利点

- Corpus mark-upと同様にコーパスの価値であり重要なもの

- 少なくとも4つの利点がある

1. 複数の意味を持つ単語を容易に区別でき（例）left、知らない言語の分析もできる
2. 再利用可能→×時間の浪費
3. 多機能性、応用可能性がある
4. 客観的指標になる
   →比較対象をする際に必要な基準になりえる

4.2 Annotationへの批判

①情報過多になるという批判

シンプルなテキストを見るべきだという主張

→ほとんどのコーパス検索ツールでは素の文も見られるため批判に値しない

4.2 Annotationへの批判

②言語学的分析、個人的解釈が必要になるという批判

→確かに必要

しかしannotationが無いことで解釈が不要になるわけではない、
客観的で精密な分析ができる点でむしろ長所である

4.2 Annotationへの批判

③コーパスの過大評価につながるという批判

①容易にアクセスできなくなる
②変更、拡大しにくくなる

①コーパス製作者はannotationに多大な労力がかかったとしても利用されることを好むため、問題にならない
むしろ著作権がアクセス規制に結びついている
②ほとんどのコーパスはサンプルコーパスで拡張の必要がない

4.2 Annotationへの批判

④正確性と一貫性に関する批判

• コーパスに注釈をつける方法は3つある
①Automatic②Computer-Assisted③Manual

コンピュータも研究者もミスをすることはある
許容できる範囲にミスを減らすことが大切

4.2 Annotationへの批判

①情報過多になるという批判
②言語学的分析、個人の解釈が必要になるという批判
③コーパスの過大評価につながるという批判
④正確性と一貫性に関する批判

→この4つの批判は無視することができる

4.3 どのようにAnnotationは達成されるか

機械や研究者によって自動、半自動、手動で行われる
自動で行われる場合、定義されたルールでコンピュータが行う

→自動の場合お金と時間がかかる

↔ 一度完成されれば素早く大量のデータ分析が可能になる
　　他で既に完成されたプログラムを踏襲することもある

【自動】

- いくつかの言語では間違えが極めて少ないためLemmatizationや品詞タグ付けが機械的に行われることが多い
- 結果が信頼できなかったり、特定の目的に対しては不正確であった場合人間による修正が必要になる

→機械の解析の曖昧な部分を人間が事後校訂すると機械のみより正確な結果になる

【手動】

- 利用できるannotationが無い場合、小さいコーパスの場合
- 高価でかかる時間が長い

ほとんどの場合、大規模なコーパスでは半自動か手動のannotation
が行われている

4.4 Annotationの種類

- 異なるレベル、様々な形式で行われている

- 音韻レベルではsyllable boundaries/phonetic anotation（音節境界/音声表記）prosodic feature（韻律）
- 形態学的レベルでは接頭辞、接尾辞、語幹
- 語彙レベルでは品詞タグ付け、レンマ、意味領域
- 統語レベルでは構文解析、treebanking,bracketing
- 談話レベルではanaphoric relations（照応関係）

4.4 Annotationの種類

- 発話や発表などの言語行為で最も一般的なannotationは品詞タグ付けである
- 多くの言語に適応されている
- いくつかのannotation(discoursal and pragmatic annotation)は発達が止まっている

・Unit4.4では言語学者によって現在使用されている一般的なannotationのタイプを紹介する

4.4.1 品詞タグ付け

- 最も一般的なAnnotationである
- 構文分析や意味注釈をするうえで基礎となる
- 同型異義語の明確化、品詞の頻度計算への応用もされている
- 多くの言語分析は品詞タグ付けに依存している
- 高度な品詞タグ付けは自動的に多くの言語に適応でき、ほとんどの調査内容に十分な精度を備えている

4.4.1 品詞タグ付け

- CLAWS
- 最も有名
- idiom listという規則に基づいた構成物によって強化されたhybrid statistical approach
- 一般的な書き言葉英語、97%の正確性を達成
- BNCコーパスを採用
- フランス語、スペイン語、ドイツ語、スウェーデン語、中国語でも発展

4.4.1 品詞タグ付け

- POSタグはそれぞれのエンコード形式で蓄積される

（例）下線

- Non mark-up aware concordance(例)MonoConc,Word Smith

→下線を使用

- Fully mark-up aware tools(例)SALA,Xaira

→SGMLやXMLが好まれる

4.4.1 品詞タグ付け

- Part-of-speech annotationでまず問題になるのは区切りである
- このプロセスはword segmentation(単語分割)やtokenization（トークン化）と言われる
- 英語のようなアルファベット言語はスペースや改行文字で区切られる
- Multiwords,mergers,variably spelled compoundsの3つを除いてorthographic正字と morpho-syntactic形態語論的な一対一の関係が規定値になっている
- CLAWSはditto taggingで複数の表現を1つの語彙単位で表す

↔sub-parts of mergersは単語を分割する

Morpho-syntactic wordをハイフンを用いたり用いなかったり、2つに分割したりしてタグ付ける

4.4.1 品詞タグ付け

- いくつかの言語、例えば中国語では文字列にスペースがなく分割が難しい
- 語彙マッチングと統計モデルの使用など複雑なプロセスが必要になる

4.4.2 Lemmatization

- 屈折語を減らし辞書形にしたannotationの一種
- （例）do, does, did, done→DO
- 語彙研究や辞書研究においてレンマ化は重要である
- 英語、フランス語、スペイン語など多くの言語で自動的にレンマ化できる
- 言語の屈折度合いによる

（例）屈折が多い言語（ロシア語）→とても有用

屈折の無い言語（中国語）→使用に限りがある

シンプルな屈折（英語）→やや不必要

4.4.3 Parsing（構文解析）

- 品詞タグ付けから派生したもの
- コーパスの文分析からコーパスの中身を分析する手続きをparsingという
- PS grammer
- Bracketing（括弧で分類すること）は句構造のラベリングに関係している
- Contraint grammer(制約文法)
- 構文分析は依存関係を包含している
- （例）主語や目的語などその語の機能的なラベリング→Treebanksに似ている

4.4.3 Parsing（構文解析）

- 品詞タグ付けの最も一般的なタイプ
- 自然言語処理（NLP）に重要
- 統語的にはparsed treebanksの方が品詞タグ付けより有用

―各単語に品詞の情報を付け加えるだけでなく構成タイプやメンバーシップを付与する

（例）構文解析を用いたほうが節の型を調べるのが容易、文法を教えることにも役立つ

4.4.3 Parsing（構文解析）

- 自動化
- 解析を自動化すると精度は品詞タグ付けに比べ、かなり低くなってしまう
- 解析コーパスは普通手動で修正を加える必要がある
- 完全に手動で作られたものもある

→コンピュータと手動を組み合わせることが一般的

4.4.3 Parsing（構文解析）

- 解析には完全解析と骨組み解析（浅い解析）がある

【完全解析】

- できるだけ詳細な統語的分析

【骨組み解析】

- 荒っぽい構成タイプが使用される
- （例）すべての名詞フレーズをNとラベルづける
- 人間による解析は骨組み解析はもちろん、完全解析でも完全に不要になることはない

4.4.4 semantic annotation(意味注釈)

- テキストに意味特徴を示すコードや単語の意味領域を割り当てる
- 少なくとも2つ大まかな意味注釈がある

①構成要素間の意味関係を印す

②テキスト内の単語の意味的特徴を印す

①semantic parsing(意味解析)としても知られる。統語レベルの注釈

②一般的なタイプ。語義タグとも呼ばれる。このタイプの注釈は内容分析にも役立つ。

（例）医者と患者の談話では医者がインタラクティブな単語を用いたほうが患者の満足度が高まる

4.4.4 semantic annotation(意味注釈)

- 基本的に知識ベースで、辞書やシソーラスなどの語彙のリソースが必要になるため品詞タグ付けよりも難しい

↔この方法が成功することもある

（例）USAS

現代英語の意味分析、21の大分類と232のサブカテゴリーからなる

まず字句単位でPOStagを付与

意味的タグ付けを行う

→92%の精度、自動化された意味注釈

4.4.5 Coreference annotation(同一指示注釈)

- 談話レベルのannotationの一種
- 懸念の一つに同一指示識別がある

（例）代名詞と名詞句間の関係性

→どのように要素が織り交ぜられ結合が達成されているのか追跡可能

→代名詞、繰り返し、置換、省略などの注釈に使用される

4.4.5 Coreference annotation(同一指示注釈)

- Lancaster/ IBM scheme
- SGML-complinant MUC scheme

- 最近まで同一指示注釈の自動または半自動のシステムで正確なものはなかった
- Xanabu（Lancaster IBMschemeを応用したもの）
- ClinkA（MUC schemeを基本としたもの）

が出てきた

4.4.6 Pragmatic annotation(実用的な注釈)

- 談話レベルの注釈の一種
- 特定の分野の音声/談話（例）医者と患者の談話
- 談話ベースのプロジェクトは世界中で行われている
- 発話タグには4種類ある

①Communicative Status(発話がわかりやすいか、完成されているか）

②Information level and status（発話の意味内容、タスクとの関係性）

③Forward-looking communicative function(その後の談話への影響)

④Backward-looking communicative function(前の談話との関係)

4.4.6 Pragmatic annotation(実用的な注釈)

- Speech Act Annotated Corpus(SPAAC)が良い例
- 電話タスクを基にした談話、41の発話カテゴリー

- Pragmatic annotationはまだ完全に自動化されていない
- 研究を助けるプログラムの開発
- XML-compliant tool
- The MATE projectなど

4.4.7 Stylistic annotation(文体注釈)

- Pragmatic Annotation
- 談話の発話行為に焦点をあてている
- Stylistic annotation
- テキストの文体注釈に焦点をあてている
- 発話と思考の表現、speech and thought presentation(S＆TP)
- Lancaster Speech, Thought and Writing Presentation Corpus(ST&WP)

→この種の唯一のコーパス

4.4.7 Stylistic annotation(文体注釈)

- 書き言葉と話し言葉の構成要素はわずかに異なる

⇔主なカテゴリーは変わらない

the direct category, the free direct category, the indirect category, the free indirect categoly

→表面的な構文は上記に書いた文体の特徴を示すことができない

自動的にこれらのカテゴリー付けをすることも難しい

Lancaster ST&WP corpusは完全に人が注釈を加えた

4.4.8 Error tagging

- エラータグ付けは学習者コーパス、言語教育に関係した特別な注釈である
- 異なる母語、背景、習熟度の学習者がよくおかす間違え
- 非母語話者の行動特徴

を調べることができる

（例）Cambridge Learner Corpus, Longman learners’ Corpus

Chinese Learner English Corpus, the JWLL(Japanese EFL Learner)など

4.4.8 Error tagging

• コーパスによってエラータグの体系は異なる
↔しかしよくあるタイプがある
Omission(省略), Addition（追加）,Misformation（誤形成）

→エラータグ付けは時間がかかる骨の折れるタスク
ルールベースや確率のプログラムは情報量の不足によって
間違えを特定することは難しい
→自動化へ向けて多くの試みがなされてきた
→いくつかのツールが発達してきた

4.4.9 Problem-oriented annotation

- 多くのannotationは広い範囲の調査質問に有効であり、多くのカテゴリーに適応可能

↔Problem-oriented annotationは異なる

①すべてのコーパスの内容ではなく特定の調査質問だけ

②具体的な研究課題との関連性

→労力があまりかからない

→時間表現や英語における同格の分析、話し方の共通点と違いの研究などに用いられてきた

4.4.9 Problem-oriented annotation

- 個々の調査や質問に依存し完全に独立したannotation
- 他の調査分析に使用することは難しい

- 調査質問にコーパスベースのアプローチをする際にとても重要なannotationである

4.5 Embedded VS. Standalone Annotation

- Annotationによって付与される情報が元のコーパスやデータに混ざってしまう

→Embedded Annotation

- SGML/XMLで表すことで元のデータとわける

→Standalone Annotation

Standalone Annotationにはより多くの利点がある

4.5 Standalone Annotationの利点

- 法に基づいた文書の供給を制御する
- 元となる文書に注釈をつけられるため変更されにくい
- 扱いにくい文書の作成を避けられる
- 異なる注釈体系を同じデータに適応することができる
- 既存の注釈、検索レベルに影響させられることなく新しい代替体系を加えることができる
- 他のレベルに影響することなく注釈のレベルを変更することができる

4.5 Standalone Annotationの利点

- 理想的で技術的に実現可能

→将来一定のタイプの注釈の標準になるかもしれない

＜問題点＞

①Corpus Annotationの複雑性によるもの

②実用性に関すること

（例）忠作のコーパス検索ツールは埋め込み注釈で作られている

Summary

- Corpus annotationの根拠について
- Annotationがどう達成されるか
- いくつかの重要なAnnotation Typeの紹介
- Standalne Corpus Annotationの利点と問題点
- →注釈はコーパスを用いた調査質問に密接に関わっている
  自動化された注釈によっては広い範囲の使用法と
  それに基づいた高いレベルの注釈がある
- このUnitでは基礎的で技術的でないcorpus annotationを扱った

ここでの探査はほんの表面的なものである